УДК 004.93

В.В. Грицюк
Центр воєнно-стратегічних досліджень, Національний університету оборони України
імені Івана Черняховського, Україна
пр. Повітрофлотський, 28, м. Київ, 03134

АЛГОРИТМ ПРОЦЕСУ АВТОМАТИЗОВАНОЇ КЛАСИФІКАЦІЇ ПОДІЙ В ІНФОРМАЦІЙНОМУ ПРОСТОРІ

V. Hrytsiuk
Center for Military Strategic Studies, National Defence University of Ukraine
named after Ivan Cherniakhovskyi, Ukraine
28, Povitroflotskyi Ave., Kyiv, 03134

ALGORITHM OF THE AUTOMATED EVENTS CLASSIFICATION PROCESS IN THE INFORMATION SPACE

У статті визначений алгоритм та деталізовано послідовні завдання для побудови ефективної моделі автоматизованої класифікації подій в інформаційному просторі. Напередодні та в ході збройної агресії Російської Федерації проти України відчутно проявилися наслідки зовнішнього негативного інформаційного впливу. Тому невідкладними є організація та здійснення протидії такому впливу. Важливою складовою цієї діяльності є класифікація (кластеризація) подій в інформаційному просторі з метою їх подальшого аналізу та формування пропозицій для прийняття рішень на протидію негативному інформаційному впливу. Враховуючи те, що в глобальному інформаційному просторі та, зокрема, в інформаційному просторі держави, в інтересах протидії такому впливу необхідно постійно обробляти значний обсяг інформації, тому вирішення завдання підвищення оперативності цього процесу передбачається за рахунок автоматизації його складових. В основу алгоритму процесу автоматизованої класифікації покладено низку виконання послідовних завдань, а саме: пошук даних, попередній відбір повідомлень (“груба” класифікація), збереження попередньо відібраних повідомлень у базі даних, визначення сукупності показників для автоматизованої класифікації інформаційних подій, попередня обробка окремого документу (індексація), розподіл повідомлень за критеріями по категоріях (“точна” класифікація), подання інформації у зручному для сприйняття вигляді (візуалізація), збереження результатів класифікації у базі даних. У запропонованому матеріалі розкритий зміст виконання цих завдань. Запропонований алгоритм слугуватиме автоматичному розподілу інформаційних подій (повідомлень) різної природи на категорії (класи) з метою підвищення оперативності оцінювання рівня негативного інформаційного впливу на цільові аудиторії для своєчасного (проактивного) реагування на його прояви.

Ключові слова: алгоритм, автоматизована класифікація, база даних, індексація, інформаційні події, негативний інформаційний вплив, терми

The article defines the algorithm and details the sequential tasks for building an effective model of automated classification of events in the information space. On the eve and during the armed aggression of the Russian Federation against Ukraine, the consequences of external negative information influence were noticeable. Therefore, the organization and implementation of counteraction to such influence is urgent. An important component of this activity is the classification (clustering) of information events in the information space in order to further analyze them and form proposals for decision-making to counteract the negative information impact. Given the fact that in the global information space and, in particular, the information space of the state in the interests of counteracting such influence, it is necessary to constantly process a significant amount of information, so the task of improving the efficiency of this process is provided by automating its components. The algorithm of the automated classification process is based on a number of consecutive tasks, namely: data retrieval, pre-selection of messages ("rough" classification), saving pre-selected messages in the database, determining a set of indicators for automated classification of information events, pre-processing a single document (indexing), distribution of messages by criteria by categories ("accurate" classification), presentation of information in a convenient form (visualization), saving the results of classification in the database. The proposed material reveals the content of these tasks. The proposed algorithm will serve to automatically divide information events (messages) of different nature into categories (classes) in order to increase the efficiency of assessing the level of negative information impact on target audiences for timely (proactive) response to its manifestations.

Keywords: algorithm, automated classification, database, indexing, information events, negative information impact, terms

Вступ

Зважаючи на значний обсяг інформації, яка потрапляє в інформаційний простір і потребує обробки в сучасних умовах та обмежений час на прийняття рішення, підвищення оперативності вирішення завдання, виявлення та оцінювання негативного інформаційного впливу на особовий склад Збройних Сил України, як необхідної умови високої результативності випереджувальних заходів протидії такому впливу, вбачається в реалізації автоматизації процедур виявлення та класифікації проявів такого впливу. Ключовим та теоретично складним для виконання цього завдання має бути процес класифікації (категоризації) інформаційних подій в інформаційному просторі з метою подальшого аналізу, оцінювання за методикою кількісного виміру та прийняття рішення щодо вживання відповідних заходів з протидії.

Аналіз літературних даних та постановка проблеми

Питання протидії негативному інформаційному впливу на різні цільові аудиторії, зокрема на особовий склад Збройних Сил України, розглядалося в працях вітчизняних науковців: В. Толубка, І. Руснака, В. Телелима, А. Рося, Т. Дзюби, Г. Пєвцова та інших [1–4]. Аналіз показує, що на сьогодні теорія протидії такому впливу обмежена на рівні концептуально декларативних положень, а тому для практики є недосконалою. У ній бракує чітких формальних методів і алгоритмів складових цього процесу, зокрема процесу класифікації інформаційних подій.

Існуючий стан системи протидії негативному інформаційному впливу на особовий склад Збройних Сил України є розбалансованим, процеси не автоматизовані. Оцінювання негативного інформаційного впливу на особовий склад Збройних Сил України та реагування на нього проводиться не інтегрально, а за окремими інформаційними проявами, причому на якісному рівні (без кількісних оцінок), що унеможливлює прогнозування ситуації та випереджувальні системні дії. Інтегральне оцінювання негативного інформаційного впливу на особовий склад Збройних Сил України здійснюється за його наслідками, через якісну оцінку рівня морально-психологічного стану особового складу Збройних Сил України на основі результатів моніторингу у військових частинах і підрозділах, відповідно діючих інструкцій [5], тобто вже після наслідків інформаційних впливів. Зазначене не дозволяє проводити випереджувальні заходи для підтримки морально-психологічного стану військ (сил), отже ефективно протидіяти такому впливу.

У цьому випадку більш доцільним було б оцінювання рівня впливу та визначення його значимості із використанням кількісної міри. Це дасть можливість успішно використати методику, описану у працях [6, 7], та реалізувати у повному обсязі наведену [8] кібернетичну модель протидії “на випередження”. Ключовим елементом методики є статистична обробка інформаційних подій та, відповідно, їх лінгвістична селекція за ознаками класифікаційної таблиці, а також “вагове” інтегрування, що визначає основну трудомісткість процесу оцінювання, від чого залежить оперативність реалізації зазначеного процесу.

Метою дослідження є розробка алгоритму процесу автоматизованої класифікації подій в інформаційному просторі для виконання функції розподілу інформаційних подій (повідомлень) різної природи на категорії (класи). Цей алгоритм автоматизації слугуватиме підвищенню оперативності загального процесу протидії негативному інформаційному впливу.

Виклад основного матеріалу

Визначимо ключовий термін дослідження.

Класифікація документів – це одне із завдань інформаційного пошуку, яке полягає у зарахуванні документа до однієї з кількох категорій на підставі його змісту [9]. Зазвичай, під класифікацією документів мається на увазі класифікація тексту, якщо не вказано інше.

На сьогодні процес класифікації інформаційних подій, зокрема в текстовому виді, можна реалізувати ручним, напівавтоматичним (автоматизованим) та автоматичним методом. При цьому розуміється, що в нашому випадку, коли вирішується питання оперативності реагування, розгляд ручного методу є недоцільним, а автоматичний метод у ряді випадків слугує для виконання складових напівавтоматичного методу, коли останній забезпечує реалізацію певного комплексного процесу.

Відповідно до [10], під автоматичною класифікацією розуміється віднесення автоматичним пристроєм об'єктів з деякої множини до того або іншого класу із заданого (скінченного) набору класів.

В [11] поняття «класифікація автоматична» еквівалентне поняттям «розпізнавання образів», «самонавчання розпізнаванню образів», «навчання без учителя» та визначається як процес автоматичного розбиття множини спостережуваних повідомлень (документів) на підмножини за вибіркою повідомлень (документів), належність яких до шуканих підмножин не вказана. Розбиття здійснюється на підставі того, як групуються повідомлення (документи) з вибірки по їх взаємній подібності або на підставі будь-яких неповних даних про шукані підмножини.

Ще досить давно В.С. Файн в тематичній статті «Енциклопедія кібернетики» [10] зазначив, що в основу автоматичної класифікації покладено аналіз інформації про кожний об'єкт, яка вводиться в пристрій. У такому випадку, інформацію про об'єкт, що класифікується, слід інтерпретувати як сукупність ознак. Тоді кожній ознаці зіставляється координата (багато градаційна або двійкова, залежно від природи ознаки) в деякому просторі ознак, де будь-який пред'явлений об'єкт буде відповідати певній точці простору. При вдалому виборі ознак точки одного класу будуть групуватися в компактні скупчення з межами, що порівняно легко апроксимуються, або в постановці ймовірності розподілами ймовірності. Поданий об'єкт, залежно від того, куди потрапляє в просторі ознак точка, що його відображає, буде автоматично класифікуватися, відповідно до прийнятого вирішального правила.

Повністю автоматичний метод класифікації передбачає набір правил або, більш загально, критеріїв прийняття рішення класифікатора, які обчислюються автоматично з навчальних даних (іншими словами, проводиться навчання машини – класифікатора). Даний підхід має нечітку кількість класів (тобто кількість класів і підкласів може змінюватись (бути гнучкою) в процесі роботи) за допомогою "машинного навчання" (Machine Learning). Але створювана множина класів може не відповідати за якістю запитам та вимогам до системи, що знижує якість обробки даних (інформації).

У нашому випадку виявлення та класифікації інформаційних подій, переважно у формі текстів, машині-класифікатору визначати типи та кількість класів не потрібно, оскільки прийнято, що вони апріорі визначені й відомі (за загальною методикою). Тому при цьому сутність процесу автоматизації не передбачає попереднього «машинного навчання», а полягає в реалізації алгоритму на основі попередньо написаних правил, відповідно до яких інформаційна подія (текст) відноситься до певного класу.

Таке рішення дозволяє забезпечити як автоматизацію процесу, так і підвищити точність класифікації, у порівнянні з «машинним навчанням».

Саме проблематиці розробки такого автоматизованого методу присвячена стаття. Для цього необхідно визначити алгоритм процесу автоматизованої класифікації подій в інформаційному просторі. Алгоритм призначений для автоматизованого розподілу інформаційних подій (повідомлень) різної природи на категорії (класи) з метою підвищення оперативності реагування на негативний інформаційний вплив на особовий склад Збройних Сил України. Алгоритм складено з ряду послідовних завдань, як представлено на рис. 1.

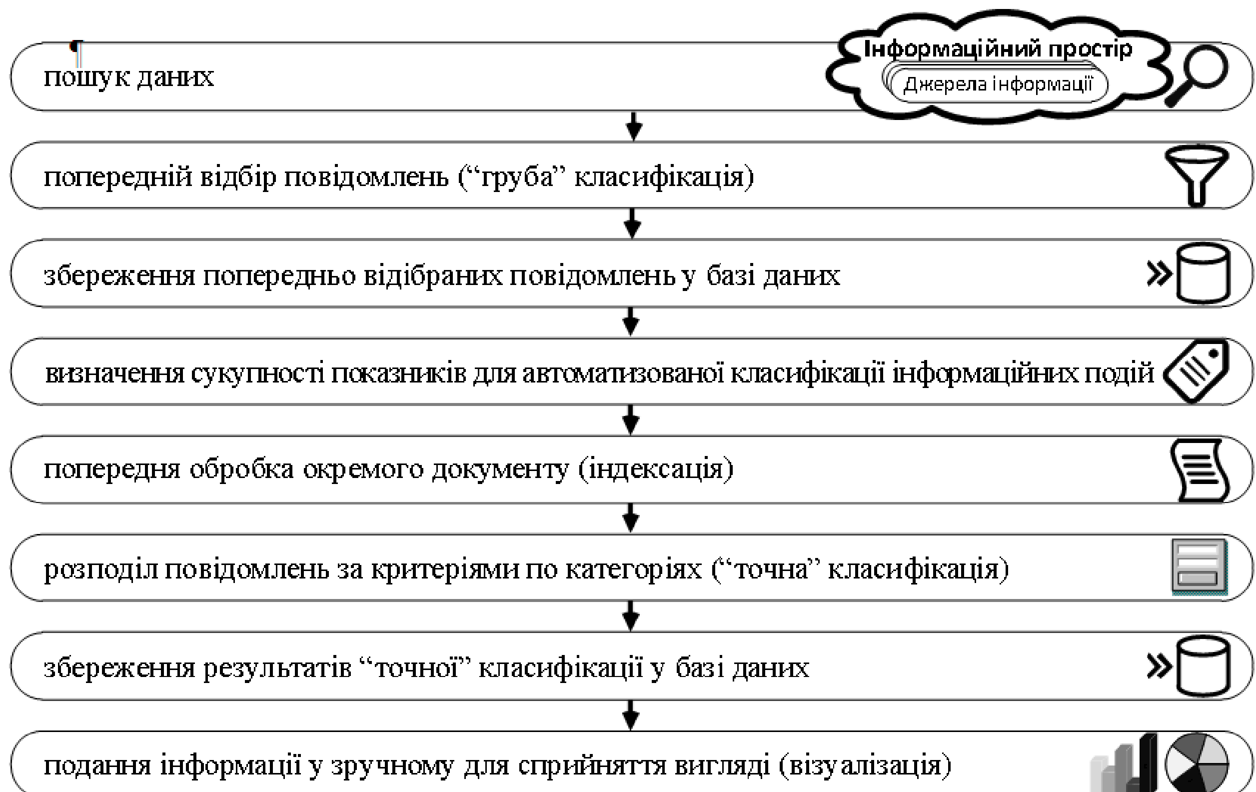


Рис. 1. Алгоритм процесу автоматизованої класифікації подій в інформаційному просторі

Опишемо кожне із завдань, посилаючись на відомі підходи у тому випадку, коли вони доцільні для реалізації схеми на рис. 1.

Пошук даних

Це інформаційний пошук неструктурованих документальних даних, зокрема, даних в документах, пошук самих документів, здобуття метаданих з документів, пошук тексту, зображень, відео та звуку у локальних реляційних базах даних, у гіпертекстових базах даних, зокрема, таких як Інтернет та локальний інтранет.

Автоматизовані системи інформаційного пошуку використовують для зменшення так званого «інформаційного перевантаження». Найвідомішим прикладом можна назвати пошукові системи в Інтернеті.

Об’єктом інформаційного пошуку є текстова інформація, зображення, аудіота відеоінформація.

Завданням інформаційного пошуку є знаходження, відповідних потребі, інформаційних об’єктів або документів серед доступного для пошуку матеріалу. Завдання для інформаційного пошуку задається у вигляді пошукового запиту (ПЗ), який може містити слова, фрази чи речення або їх комбінацію. Переважна більшість пошукових систем орієнтована на роботу з пошуковими термінами (термами) – словами або словосполученнями, які пошукова система розпізнає як одне ціле [12].

Попередній відбір повідомлень («груба класифікація»)

Загалом, результати інформаційного пошуку повинні відповідати таким вимогам [12]:

– релевантність – стосується результатів роботи пошукової системи й експертної системи; ступінь відповідності запиту й знайденого, тобто доречність результату. Одне з найбільш близьких поняттю «релевантності» – «адекватність», тобто оцінка ступеня відповідності практичної та соціальної застосовності результату варіантів вирішення завдання;

– пертинентність – співвідношення обсягу корисної інформації до загального обсягу отриманої інформації.

Завдання «попереднього відбору повідомлень» являє собою, по суті, 1-й «грубий» етап відбору («грубе сито»). Тобто, розглядається певна ситуація, коли відбираються необхідні повідомлення про інформаційні події із всієї множини інформаційного простору. Іншими словами, цей процес можна вважати «грубим ситом» подій в інформаційному просторі або «грубою класифікацією» інформаційних подій. Процес виконується за ключовими словами, що відносяться до певної окремої категорії. У такому випадку, класифікацію можна також назвати категоризацією.

Для здійснення попереднього інформаційного пошуку потрібно мати доступ до збірки (обсягу) інформаційних об'єктів (бібліотеки, бази даних тощо) і автоматизовану систему (сервіс або програму), яка здійснює пошук. При цьому попередній відбір повідомлень передбачає включення до масиву усіх повідомлень, які за ознаками можна віднести до єдиного класу (категорії), яка, у нашому випадку, характеризує ці повідомлення як такі, що негативно впливають на особовий склад Збройних Сил України, тобто пошуковий запит має узагальнювати ознаки одразу усіх класів інформаційних повідомлень (відповідно до методики [6, 7]). Отже має бути визначено певний мета-тег для використання в ПЗ. Пошукова система переглядає всі доступні інформаційні одиниці (документи) зі збірки й відбирає відповідні до мета-тегу ПЗ. Результатом пошукової роботи є упорядкований список документів, який укладається, згідно з певним принципом. Таким чином, процес попереднього інформаційного пошуку – це алгоритм, який, переглядаючи доступну збірку інформаційних об'єктів за певний проміжок часу, формує попередній набір документів: список $D = \{d_i\}, i = \overline{1,n}$, відповідно до ПЗ.

Для процедури формування ПЗ необхідно здійснити формалізацію основних понять, сукупно для усіх класів інформаційних повідомлень, та створити мета-тег загального значення.

З цього приводу слід зазначити, що характеристикою певного веб-ресурсу являються дві основоположні складові, це: метадані та внутрішнє смислове навантаження, тобто власне основний текст документа. Ці складові використовуються в тому числі для того, щоб допомогти пошуковим машинам віднести веб-сторінку (інформацію) до тієї чи іншої тематичної сукупності. Тобто пошукові системи порівнюють (ототожнюють) ПЗ, який може бути представлений мета-тегом або тегами (сукупністю ключових слів) з метаданими та власне текстом вебресурсу (документу). Тому спочатку необхідно визначити теги для кожного з класів. Сукупність всіх цих тегів про кожен клас об'єднується у загальний мета-тег, який і буде характеризувати усю сукупність класів. Відповідно, на першому етапі класифікації інформаційних подій, який ще, іншим чином, можна назвати "грубим ситом", цей мета-тег буде використаний для формування ПЗ та подальшого встановлення релевантності з метаданими та текстом веб-ресурсів. Згідно із цим правилом, формується множина документів $D = \{d_i\}$ для подальшої класифікації (категоріювання).

Збереження попередньо відібраних повідомлень у базі даних

Після відбору повідомлень (документів) необхідно зберегти їх у базі даних у початковому оригінальному вигляді, зокрема для подальшої перевірки (аналізу), для здійснення класифікації повідомлень. Цей етап є резервуванням отриманих даних, у відповідності до ПЗ.

Визначення сукупності показників для автоматизованої класифікації інформаційних подій

Процес класифікації повідомлень масиву $D = \{d_i\}$ за відомими класами може проводитись за, так званими, "частковими класифікаторами". Такий класифікатор є переліком категорій аналізу, індикаторів (прийнятих одиниць реєстрації), основою алгоритму наступних дій. Від його вибору залежить якість процесу автоматизованої класифікації.

Категорії аналізу – ключові елементи дослідницької концепції, значеннєві одиниці, які реєструють, відповідно до поставленої мети. Список категорій повинен бути вичерпним, забезпечувати можливість однозначного співвіднесення частин тексту з конкретною категорією (класом). У нашому випадку список категорій є відомим, відповідно до розділу 1, цей список включає 22 категорії або класи (за іншими методиками це число може бути іншим).

Індикатори – ознаки вираження певної сутності тексту, які є його частинами, що характеризують належність повідомлення (тексту) до окремої категорії (значеннєвої одиниці). Ними можуть бути символи, слова, терміни, словосполучення, ситуації, судження, репліки, інтонації, які дають змогу визначити роль у тексті кожної категорії. Вона може виражатися у тексті по-різному: від окремих символів чи слів до суджень або абзаців.

Попередня обробка окремого документа (індексація)

Для вирішення задачі автоматичної класифікації текстів, в першу чергу необхідно виконати попередню обробку документів з множини $D = \{d_i\}, i = \overline{1,n}$ інформаційних об’єктів, яку називають індексацією. На цьому етапі документи, що мають вигляд послідовності символів, перетворюються до виду, придатного для машинних алгоритмів, у відповідності до задачі класифікації. Зазвичай, за допомогою алгоритмів реалізації цієї функції, опрацьовуються вектори в так званому просторі ознак [13].

Індексацію можна представити у вигляді трьох етапів, як зображено на рис. 2 [14].

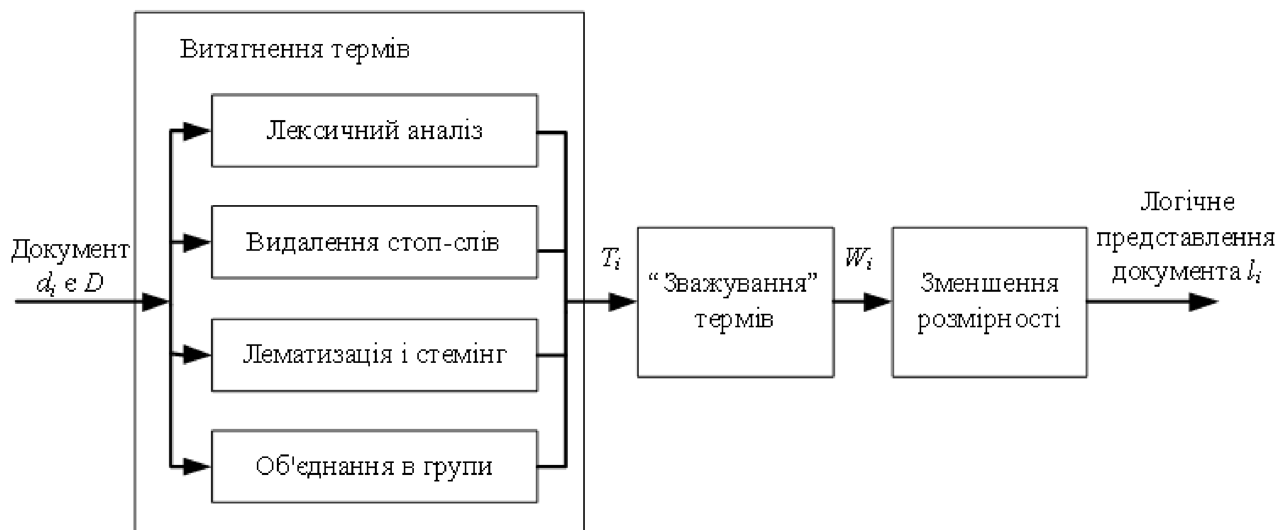


Рис. 2. Попередня обробка (індексація) документа

На виході усієї процедури індексації отримується логічне (формальне) представлення документа для подальшої обробки. Розглянемо елементи процедури індексації.

Витягнення термів

Витягнення термів, або витягнення ознак – це процес розбиття тексту на простіші об’єкти, які також називаються термами. Результат цього процесу – це множина термів T_i, які використовуються для отримання вагових характеристик документа. Процес передбачає ряд процедур, що наводяться.

Лексичний аналіз. Лексичний аналіз – перший крок вилучення термів. На цьому етапі відсіюються всі символи, які не є буквами (наприклад, розділові знаки й html-теги).

Видалення стоп-слів. Стоп-слова – це слова, що не несуть будь-якого самостійного смислового навантаження. До стоп-слів належать прийменники, сполучники й займенники [15]. З метою зменшен-

ня розмірності простору термів, індексатор не враховує стоп-слів і видаляє їх при аналізі. Так само стоп-слова сильно впливають на відбір ключових слів. Якщо їх не видалити, вони засмічують безліч термів, так як зустрічаються часто в тексті.

Додатково варто відзначити, що бувають ситуації, коли є сенс відмовитися від видалення стоп-слів. У статті [16] показано, що від тематики до тематики сильно змінюються фразові дієслова й точність класифікації, з урахуванням стоп-слів, може зрости.

Лематизація й стемінг. Лематизація – це приведення кожного слова в документі до його нормальної форми. Зокрема, в українській мові нормальними формами вважаються:

- для іменників й прикметників – називний відмінок, однина, чоловічий рід;
- для дієслів, дієприкметників й дієприслівників – дієслово в невизначеній формі.

При побудові безлічі термів часто нехтують формами слова. Це виправдано, так як, зберігаючи форми слів, простір термів і структура зберігання будуть швидко рости, що погіршить продуктивність, а статистика буде ділитися між формами одного слова, погіршуючи загальну картину.

Стемінг – відкидання змінюваних частин слів, головним чином закінчень. Ця технологія простіша, не вимагає зберігання словника слів або великого набору правил. Технологія заснована на правилах морфології мови. Недолік стемінгу – велике число помилок. Стемінг добре підходить, наприклад, для англійської мови, але гірше – для української.

Одна з проблем при розгляді слів в якості термів – це їх семантична неоднозначність, яку умовно можна поділити на дві групи:

1. Синоніми – слова однієї частини мови, різні за звучанням і описом, але мають схоже лексичне значення (йти – крокувати, сміливий - хоробрий);
2. Омоніми – різні за значенням, але однакові за написанням одиниці мови (міна – вираз обличчя або вибуховий снаряд).

Вирішити цю невизначеність можна, використовуючи контекст слова в реченні. Для цього використовуються методи морфологічного й лінгвістичного аналізу [17].

Об'єднання в групи (N-грами). Об'єднання в групи – це процес об'єднання декількох послідовних слів в одну групу, яку називають N-грамою. У такому випадку, кожна N-грама розглядається як самостійний терм $t_{ji} \in T_i$ документа.

Якщо розділити текст на кілька невеликих фрагментів, представлених N-грамами, їх легко порівняти одна з одною і, таким чином, отримати ступінь подібності контрольованих документів, що, зокрема, часто застосовується у виявленні плагіату. Використовуючи N-грами, також можна ефективно знайти кандидатів для заміни слів з помилками правопису. Основний недолік застосування N-грам, це швидко зростаючий обсяг пам'яті, необхідний для їх зберігання.

Текстова інформація документа $d_i \in D$ подається як множина термів $T_i = \{t_{1i}, \dots, t_{m_i i}\}$. Кожному терму $t_{ji} \in T_i$, $j = \overline{1, m_i}$, ставиться у відповідність деяка "вага" w_{ji}. Ця функція є числовою характеристикою розповсюдженості цього слова в документі $d_i \in D$. При цьому враховується не тільки частота повторюваності слова в тексті, а також інші ознаки, такі як: порядок слів, повторюваність у заголовку, слово, що міститься в метаданих джерела інформації та інші. На підставі цих ознак, кожному терму в тексті відповідає його "вага".

Попередня обробка документа, в такому випадку, це перетворення послідовності термів документа в m-вимірний векторний простір. Процес отримання вектора "ваг" для документа називається індексацією документа.

Зважування термів з використанням статистичної обробки

Один з відомих методів представити "вагу" терму – метод TF-IDF.

TF (term frequency – частота терму) – відношення числа входження деякого терму до загальної кількості термів документа. Так оцінюється домінування окремого терму в межах документа [18].

Нехай $v\{t_{ji}\}$ – число входжень терму t_{ji} в документ $d_i \in D$.

Тоді частота терму t_{ji} визначається

$$TF\{t_{ji}\} = \frac{v\{t_{ji}\}}{\sum_{j=1}^{m_i} v\{t_{ji}\}}, i = \overline{1,\mathrm{n}}, j = \overline{1,\mathrm{m_i}} \quad (1)$$

IDF (inverse document frequency – зворотна частота документа) – інверсія частоти, з якою деякий терм зустрічається в усіх документах множини $D = \{d_i\}$. Врахування IDF зменшує "вагу" широковживаних термів. Для кожного унікального терму в межах конкретної множини документів існує тільки одне значення IDF [18].

$$\mathrm{IDF}(t_{ji}) = \frac{\mathrm{n}}{\sum_{i=1}^{n}(d_i : t_{ji} \in d_i)}, \quad d_i \in D, \quad (2)$$

де n – кількість документів у множині $D = \{d_i\}$;

$\sum_{i=1}^{n}(d_i : t_{ji} \in d_i) \geq 1$ – кількість документів, в яких зустрічається терм t_{ji}.

TF-IDF – статистична міра, яка використовується для оцінки важливості терму в контексті документа, що є частиною множини D. Відповідно до [13, 18], "вага" деякого терму пропорційна кількості вживання цього слова в документі $d_i \in D$ і обернено пропорційна частоті вживання слова в інших документах множини D:

$$w_{\mathrm{ji}} = \mathrm{TF}\{t_{ji}\} \times \mathrm{IDF}(t_{ji}), \ j = \overline{1,\mathrm{m_i}} \quad (3)$$

У результаті процесу "зважування термів" отримується вагова характеристика кожного документа $d_i \in D$, як кортеж "ваг" термів.

$$W_i = \{w_{1i}, \ldots, w_{m_i i}\}, i = \overline{1,\mathrm{n}} \quad (4)$$

Реалізація алгоритмів TF, TF-IDF вже існує в бібліотеках для роботи з текстами та виконується на мові програмування python. Для прискорення роботи з великими матрицями термів використовується бібліотека numpy [14].

Зменшення розмірності векторів

Для скорочення розмірності векторів можна не враховувати рідкісні слова, які збільшують розмір простору, але, як правило, не несуть корисної для класифікатора інформації. Також можна не розглядати слова, що часто зустрічаються, такі як артиклі тощо. Для кожного терму можна визначити його коефіцієнт значущості, тобто наскільки цей терм корисний для класифікації. Цю характеристику можна визначити, грунтуючись на кореляції між частотою появи слова в документі й приналежністю цього документа до однієї або декількох категорій.

Крім видалення зайвих термів, можна групувати кілька термів в один. Наприклад, можна групувати разом синоніми. Ще один підхід – "спільна зустрічальність" (cooccurrence): об'єднувати слова, які часто зустрічаються в одному оточенні. Наприклад, в словосполученнях «керівник компанії», «директор компанії» слова «керівник» та «директор» зустрічаються перед словом «компанія». Тому їх можна об'єднати в один штучний терм. У загальному випадку, для слів визначається якась метрика близькості, й групи близьких слів склеюються в один терм. Вага такого терму в кожному конкретному документі розраховується з ваг представників групи, які зустрічаються в цьому документі.

Таким чином, логічне представлення документа $d_i \in D$ в такому випадку отримується виокремленням всіх значущих термів і визначенням їхньої "ваги". Після процесу функції "зменшення розмірності" в кожному документі d_i отримуємо кількість термів $j = \overline{1,k}$. При цьому розуміється

$$k < m,$$

де k – кількість термів до моменту "зменшення розмірності векторів"; m – кількість термів після зазначеної функції.

У підсумку, кожен документ повинен бути представлений вектором k-вимірної розмірності $\overline{d_i} = \langle w_{1i}, \ldots, w_{k_i i} \rangle$, де кожен компонент w_{ji} є вагою *j*-го терму з множини термів *T* в документі d_i. Отриманий в результаті *n*-вимірний простір векторів прийнято називати *простором ознак* для документів множини *D*. Кожен індек-

сований документ $d_i \in D$ в результаті обробки подається в вигляді логічного представлення [15].

$$l_i = \{w_{1i}, \dots, w_{k_i i}\}, i = \overline{1,n}. \quad (5)$$

У подальшому, логічне представлення документа l_i буде тією ознакою документа, за якою буде проводитись автоматична класифікація документа, тобто віднесення до тієї чи іншої категорії. Логічне представлення документа є, по суті, набором "ваг" термів. Ці "ваги" будуть ранжировані. І найбільш пріоритетні з них порівнюються (ототожнюються) з тегами наперед відомих категорій (класів) за спеціальними правилами. Найбільш релевантні співпадіння "ваг" термів з тегами категорій будуть підставою для віднесення до тієї чи іншої категорії.

Розподіл повідомлень за критеріями по категоріях ("точна" класифікація)

Після здійснення індексації наступним йде етап (функція) класифікації (категоризації).

Схематично процес класифікації (категоризації) інформаційних подій зображено на рис. 3.

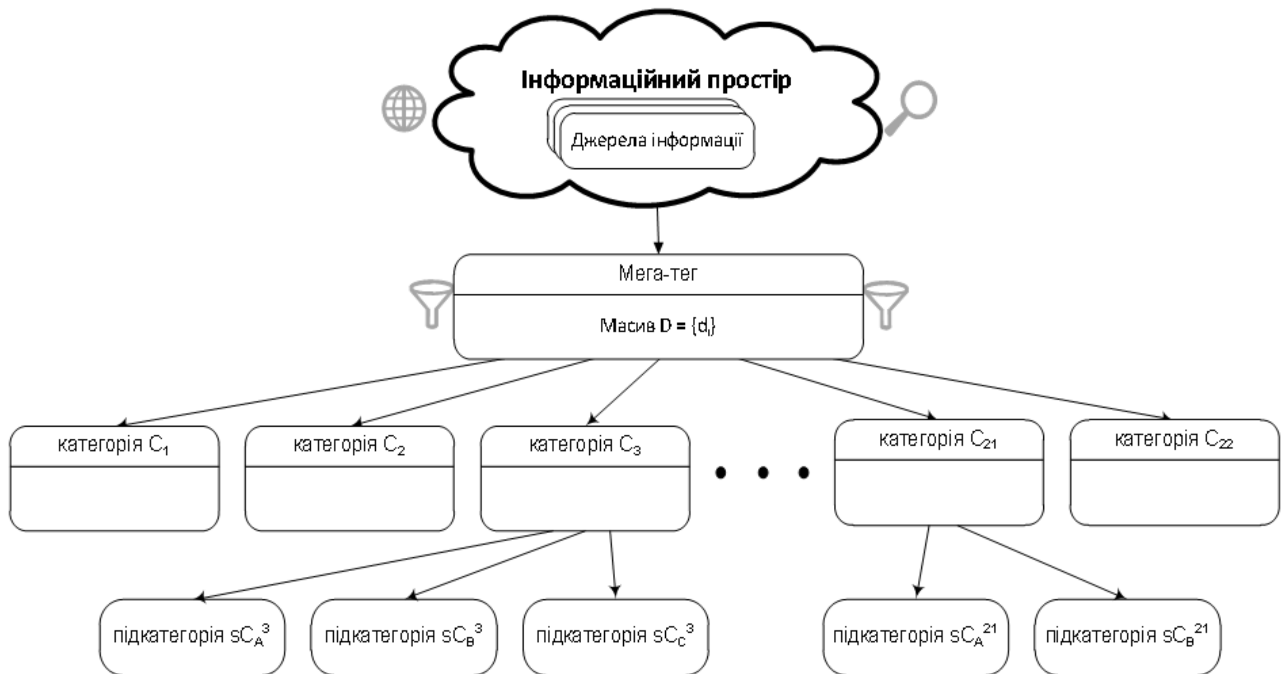


Рис. 3. Ієрархічне представлення процесу класифікації (категоризації)

Критерій категорії – сформульована умова включення того чи іншого документа в окрему категорію (клас). Критерій встановлює співвідношення (семантичний зв'язок) між темою категорії й документом, що включається в неї. При цьому елемент, що включається, є граматичним суб'єктом висловлювання про відношення, а тема категорії – складовою частиною предиката висловлювання про відношення.

Критерії категоризації документів. У категорію включаються документи або підкатегорії, відповідні встановленим критеріям цієї категорії. Основні вимоги до критерію будь-якої категорії для документа – істотність ознак категоризації, його унікальність, відповідність базовим правилам (перевіряємість, нейтральна точка зору тощо).

Істотність ознаки означає наявність зафіксованого в авторитетних джерелах способу класифікації об'єктів, що підпадають під класифікацію (категоризацію). Цей спосіб повинен бути відповідним до критерію категорії. Під *унікальністю* розуміється неприпустимість створення множинних категорій з ідентичними або дуже близькими критеріями включен-

ня. *Верифікованість* означає можливість перевірити виконання критерію категорії для кожної зі статей на підставі авторитетних джерел. Нейтральна точка зору вимагає дотримання нейтральності в формулюванні критерію.

Подання інформації в зручному для сприйняття вигляді (візуалізація)

На практиці результати аналізу контенту найчастіше представляються рядами діаграм: стовпчастих чи кругових. Також, для відображення відносин між одиницями аналізу контенту та результатів їх категоризації використовуються такі стандартні засоби відображення структур, як різні графи. Візуалізація відбувається за допомогою деяких комп'ютерних програм. Наприклад, Microsoft Excel та SPSS. Презентувати дані допомагають програми на кшталт Microsoft PowerPoint та Prezi.

Збереження результатів класифікації у базі даних

Всі розрахунки та візуалізовані результати класифікації (категоризації) зберігаються у базах даних або на матеріальних носіях інформації. Ці дані будуть далі враховані за допомогою методики виявлення та оцінювання негативного інформаційно-психологічного впливу на особовий склад Збройних Сил України, яка описана в першому розділі. Після повного комплексу зазначених процедур інформація буде надаватися особам, що приймають рішення, на подальшу протидію такому впливу.

Висновки

В основу автоматичної класифікації доцільно покласти аналіз інформації про кожний об'єкт, яким є зафіксоване повідомлення в інформаційному просторі. В такому випадку, інформацію про об'єкт, що класифікується, слід інтерпретувати як сукупність ознак. Пред'явлений об'єкт класифікувати відповідно до прийнятого вирішального правила.

Запропонований автоматизований підхід полягає в написанні правил, згідно яких автоматичним методом можна зарахувати текст до тієї чи іншої категорії.

Розроблений алгоритм слугуватиме автоматичному розподілу інформаційних подій (повідомлень) різної природи на категорії (класи) з метою підвищення оперативності оцінювання рівня негативного інформаційного впливу на особовий склад Збройних Сил України для своєчасного (проактивного) реагування на його прояви.

Література

1. Толубко В.Б. Концептуальні основи інформаційної безпеки України / В.Б. Толубко, С.Я. Жук, В.О. Косевцов // Наука і оборона. – 2004. – № 2. – С. 19-25.
2. Руснак І.С. Розвиток форм і способів ведення інформаційної боротьби на сучасному етапі / І.С. Руснак, В.М. Телелим // Наука і оборона. – 2000. – № 2. – С. 18-23.
3. Основи стратегії національної безпеки та оборони держави: підруч. / О.П. Дузь-Крятченко, Т.М. Дзюба, А.О. Рось, ін. – 2-ге вид., доп. і випр. – К.: НУОУ, 2010. – 591 с.
4. Інформаційно-психологічна боротьба у воєнній сфері: монографія / Г.В. Пєвцов, А.М. Гордієнко, С.В. Залкін, С.О. Сідченко, А.О. Феклістов, К.І. Хударковський. – Х.: Вид. Рожко С.Г., 2017. – 276 с.
5. Інструкція про порядок оцінки морально-психологічного стану в Міністерстві оборони України та Збройних Силах України (затверджено наказом МО України від 21.05.2013 № 335, зі змінами, внесеними наказом МО України від 17.12.2015 № 728, зареєстровано в Мін'юсті України 11.01.2016 № 29/28159).
6. Методичний підхід до виявлення та оцінювання негативного інформаційно-психологічного впливу на особовий склад військ (сил) / П.М. Сніцаренко, Ю.О. Саричев, Ю.І. Міхєєв, М.В. Праута // Наука і оборона. – № 3-4. – 2017. – С.18-25.
7. Підсистема моніторингу інформаційного простору як необхідна складова системи протидії негативному інформаційно-психологічному впливу на особовий склад Збройних Сил України / П.М. Сніцаренко, Ю.О. Саричев, В.А. Ткаченко, О.А. Мотузяник // Наука і оборона. – № 1. – 2018. – С.29-33.
8. Аналіз стану виявлення та оцінки негативного інформаційного впливу на особовий склад ЗС України в системі протидії такому впливу / В.В. Грицюк // Збірник наукових праць ЦВСД НУОУ. – № 2(66). – 2019. – С.52-61.
9. Christopher D. Manning, Hinrich Schütze An Introduction to Information Retrieval Draft. Online edition. Cambridge University Press. – 2009. – 544 p.
10. Енциклопедія кібернетики : у 2 т. / за ред. В.М. Глушкова. — Київ : Гол. ред. Української радянської енциклопедії, 1973. – Т.1. – С.490.
11. Словарь по кибернетике: Св. 2000 ст. /Под ред.

В.С. Михалевича. – 2-е изд. – К.: Гл. ред. УСЭ им. М. П. Бажана, 1989. – 751 с.
12. Ланде Д.В., Снарский А.А., Безсуднов И.В. Интернетика: Навигация в сложных сетях: модели и алгоритмы. — М.: Либроком (Editorial URSS), 2009. – 264 с.
13. Великий тлумачний словник сучасної української мови / уклад. та голов. ред. В.Т. Бусел. – Київ; Ірпінь: Перун, 2005. – VIII, 1728 с.
14. Агеев М.С. Методы автоматической рубрикации текстов, основанные на машинном обучении знаниях экспертов, 2004. – С. 6.
15. Попков М.И. Автоматическая система классификации текстов для базы знаний предприятия // International Journal of Open Information Technologies ISSN: 2307-8162 vol. 2, no. 7, 2014, P. 1–10.
16. «Глоссарий» // Веб-студия pawlov.info (http://www.pawlov.info/index.php/glossarij)
17. Ellen Riloff, «Little Words Can Make a Big Difference for Text Classification» // Department of Computer Science, University of Utah
18. Воронцов К.В. «Вероятностное тематическое моделирование» // Лекции по Машинному обучению, Октябрь 2013.
19. Губин М.В. «Модели и методы представления текстового документа в системах информационного поиска», 2005. С. 11-12.
20. Rehman A., Haroon A., Saeed M., Feature Extraction for Classification of Text Documents, 2012. – P. 233-235.

References

1. Tolubko V.B. Konceptualni osnovy informacijnoyi bezpeky Ukrayiny / V.B. Tolubko, S.Ya. Zhuk, V.O. Kosevczov // Nauka i oborona. – 2004. – № 2. – P. 19-25.
2. Rusnak I.S. Rozvytok form i sposobiv vedennya informacijnoyi borot`by na suchasnomu etapi / I.S. Rusnak, V.M. Telelym // Nauka i oborona. – 2000. – № 2. – P. 18-23.
3. Osnovy strategiyi nacional`noyi bezpeky ta oborony derzhavy: pidruch. / O.P. Duz`-Kryatchenko, T.M. Dzyuba, A.O. Ros`, in. – 2-ge vyd., dop. i vypr. – K.: NUOU, 2010. – 591 p.
4. Informacijno-psyxologichna borot`ba u voyennij sferi: monografiya / G.V. Pyevczov, A.M. Gordiyenko, S.V. Zalkin, S.O. Sidchenko, A.O. Feklistov, K.I. Xudarkovs`ky`j. – X.: Vyd. Rozhko S.G., 2017. – 276 p.
5. Instrukciya pro poryadok ocinky moral`no-psyxologichnogo stanu v Ministerstvi oborony Ukrayiny ta Zbrojnyx Sylax Ukrayiny (zatverdzheno nakazom MO Ukrayiny vid 21.05.2013 № 335, zi zminamy, vnesenymy nakazom MO Ukrayiny` vid 17.12.2015 № 728, zareyestrovano v Minyusti Ukrayiny 11.01.2016 № 29/28159).
6. Metodychnyj pidxid do vyyavlennya ta ocinyuvannya negatyvnogo informacijno-psyxologichnogo vplyvu na osobovyj sklad vijsk (syl) / P.M. Sniczarenko, Yu.O. Sarychev, Yu.I. Mixyeyev, M.V. Prauta // Nauka i oborona. – № 3-4. – 2017. – P.18-25.
7. Pidsystema monitoryngu informacijnogo prostoru yak neobxidna skladova systemy protydiyi negatyvnomu informacijno-psyxologichnomu vplyvu na osobovyj sklad Zbrojnyx Syl Ukrayiny / P.M. Sniczarenko, Yu.O. Sarychev, V.A. Tkachenko, O.A. Motuzyany`k // Nauka i oborona. – № 1. – 2018. – P.29-33.
8. Analiz stanu vyavlennya ta ocinky negatyvnogo informacijnogo vplyvu na osobovyj sklad ZS Ukrayiny v systemi protydiyi takomu vplyvu / V.V. Grycyuk // Zbirnyk naukovyx pracz` CzVSD NUOU. – № 2(66). – 2019. – P.52-61.
9. Christopher D. Manning, Hinrich Schütze An Introduction to Information Retrieval Draft. Online edition. Cambridge University Press. – 2009. –544 p.
10. Encyklopediya kibernetyky : u 2 t. / za red. V.M. Glushkova. — Kyyiv : Gol. red. Ukrayins`koyi radyans`koyi encyklopediyi, 1973. – T.1. – P.490.
11. Slovar` po ki`bernetike: Sv. 2000 st. /Pod red. V.S. Mixalevicha. – 2-e izd. – K.: Gl. red. USE im. M. P. Bazhana, 1989. – 751 p.
12. Lande D.V., Snarskij A.A., Bezsudnov I.V. Internetika: Navigaciya v slozhnyx setyax: modeli i algoritmy. — M.: Librokom (Editorial URSS), 2009. – 264 p.
13. Velykyj tlumachnyj slovnyk suchasnoyi ukrayins`koyi movy / uklad. ta golov. red. V.T. Busel. – Kyyiv; Irpin`: Perun, 2005. – VIII, 1728 p.
14. Ageev M.S. Metodы avtomaty`cheskoj rubry`kacy`y` tekstov, osnovannыe na mashy`nnom obucheny`y` znany`yax эkspertov, 2004. – P. 6.
15. Popkov M.I. Avtomaticheskaya sistema klassifikacii tekstov dlya bazy znanij predpriyatiya // International Journal of Open Information Technologies ISSN: 2307-8162 vol. 2, no. 7, 2014, P. 1–10.
16. «Glossarij» // Veb-studiya pawlov.info (http://www.pawlov.info/index.php/glossarij)
17. Ellen Riloff, «Little Words Can Make a Big Difference for Text Classification» // Department of Computer Science, University of Utah
18. Voronczov K.V. «Veroyatnostnoe tematicheskoe modelirovanie» // Lekcii po Mashinnomu obucheniyu, Oktyabr` 2013.
19. Gubin M.V. «Modeli i metody predstavleniya tekstovogo dokumenta v sistemax informacionnogo poiska», 2005. P. 11-12.
20. Rehman A., Haroon A., Saeed M., Feature Extraction for Classification of Text Documents, 2012. – P. 233-235.

Надійшла до редакції 10.01.2020